ProxyGenBench: Benchmarking Proxy-Guided Video Generation

Zhenchen Tang[1,4,5,*], Haolin Lu[2,4,*], Songlin Yang[3,4], Xiaotong Zhao[4], Ran Zhang[4], Bo Peng[1,5], Shuai Li[4], Haotian Fan[4], Haobin Zhong[4], Yanni Zhang[4], Songlian Li[4], Ning Zhang[4], Xun Gao[4], Alan Zhao[4], Jing Dong[1,5,†]

[1]New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

[2]University of California, San Diego [3]The Hong Kong University of Science and Technology

[4]Tencent [5]School of Artificial Intelligence, University of Chinese Academy of Sciences

[*]Equal contribution. [†]Corresponding author.

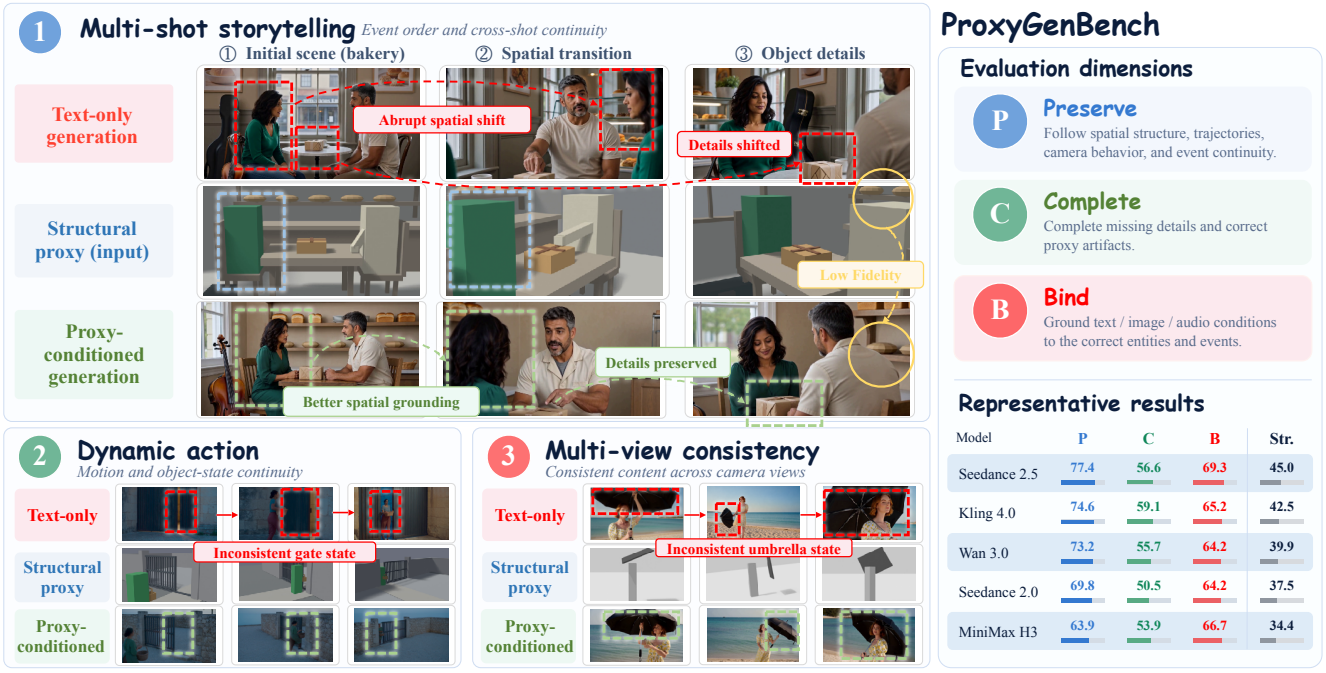


Figure 1. **Overview of ProxyGenBench.** The BAKERY example shows how proxy guidance improves spatial and object-detail consistency across shots, while the DYNAMICACTION and MULTIVIEW examples highlight consistency in gate and umbrella states. The proxies supply coarse scene structure rather than complete target videos. The right panel summarizes *Preserve*, *Complete*, and *Bind* dimensions and reports rounded scores from Tab. 3(a) on a 0–100 scale; Str. denotes our *Strict* score, a structure-aware aggregation of all three dimensions.

Abstract

Recent video generators produce realistic videos, yet controlling how scenes unfold remains challenging. Structural proxies, such as animated 3D blockouts, guide generation by specifying scene layout, motion, camera behavior, and event order. However, these coarse inputs leave details unspecified and may contain artifacts that should not be reproduced. Successful proxy-guided generation must therefore ***Preserve*** *intended structure,* ***Complete*** *missing content and correct proxy artifacts, and* ***Bind*** *text, image, and audio conditions to the right entities and events. We introduce* ***ProxyGen-***

Bench *to study these interacting requirements, with 312 evaluation cases from 60 scenes across three scene tasks and controlled variations in proxy fidelity, text detail, and multimodal inputs. We evaluate five video generators using verified, case-specific checklists. Experiments show that proxies improve spatiotemporal control, but greater proxy detail can coincide with poorer completion, while strong structural adherence does not ensure correct binding.*

1. Introduction

Video generation has advanced rapidly in realism, but precise control over how a scene unfolds remains challenging [6, 21, 31]. Generated videos may contain incorrect interactions, inconsistent spatial relationships, or discontinuities across shots, even when individual frames appear plausible [2, 23]. Text prompts and reference images provide semantic and appearance guidance, but often leave object trajectories, camera behavior, and event timing implicit [12, 28]. Structural proxies, such as animated 3D blockouts, complement these inputs by representing scene layout, motion, camera behavior, and event order in a coarse video sequence.

While structural proxies offer more explicit control over scene organization, their guidance is not without limitations, as illustrated in Fig. 1. In the *bakery* scene, text-only generation introduces an abrupt spatial shift and changes object details across shots, while proxy guidance helps preserve scene layout and object continuity. Similar improvements are observed in the *dynamic-action* and *multi-view* examples. However, proxy guidance often contains low-fidelity geometry and motion that may mislead the generator and hinder realistic scene realization, as illustrated by the low-poly donuts in the *bakery* example. What should be preserved from the proxy and what should be refined ultimately depends on the user's intent, making additional conditions, such as text and reference images, essential for guiding the generation. This raises a central question: *how can generators preserve useful proxy constraints without inheriting their artifacts or overlooking information from other conditions?* Existing benchmarks provide limited insight into how these requirements interact in proxy-guided generation.

To address this gap, we identify three complementary abilities essential for successful proxy-guided generation: **Preserve** intended structure, **Complete** missing content and correct proxy artifacts, and **Bind** additional conditions to the appropriate entities and events. To evaluate these abilities, we introduce **ProxyGenBench**, comprising 312 evaluation cases from 60 scenes across dynamic action, multi-shot storytelling, and multi-view consistency challenges.

Our scenes cover four potential applications of proxy-guided generation: enhancing reference-to-video workflows, enabling code-driven proxy authoring, supporting game-like world modeling, and assisting 3D animation completion. For each scene, we provide multiple variants to study how models respond to different levels of conditioning detail and balance guidance across modalities. Low-fidelity proxies leave more room for refinement but may provide insufficient structural guidance, while high-fidelity proxies are easier to follow but risk transferring unwanted details. Similarly, simple text prompts may underspecify user intent, whereas detailed prompts or additional image references may distract models from the proxy. Through matched and targeted configurations, we examine when proxies improve control, how proxy fidelity affects completion, and whether additional conditions are correctly integrated into the generated scene.

To evaluate model performance across these variations, we organize 32 fine-grained criteria into eight capabilities under Preserve, Complete, and Bind, and instantiate applicable criteria as verified, case-specific checklist questions. Each checklist is grounded in the inputs available to the generator and fixed across models for the same case. VLM judgments use visual or audio evidence to assess structural adherence, completion, and binding beyond proxy resemblance. We report the three dimension scores alongside a structure-sensitive *Strict* score, which gives additional emphasis to critical Preserve constraints because structural control is the basic requirement of proxy-guided generation.

Experiments on five video generators show that proxy guidance improves spatiotemporal control, particularly action transitions and cross-shot spatial consistency. However, a high average Preserve score does not guarantee that all critical structural constraints are satisfied together. In scene-matched comparisons, moving from coarse to detailed proxies increases Preserve but lowers Complete for all five models, revealing a preservation–completion tension. Results on Studio 3D proxies also show that strong structural adherence can coexist with incorrect condition binding. Together, these findings highlight the need to preserve intended structure selectively, correct proxy artifacts, and ground supplementary information within the same generated scene.

Our contributions are threefold:

- **Benchmark.** We build controlled proxy-guided generation cases varying proxy detail and text, image, and audio requirements.
- **Evaluation.** We develop verified, case-specific checklists to assess intended structural constraints, completion of unfinished content, correction of proxy artifacts, and grounding of additional conditions.
- **Analysis.** We quantify proxy benefits and characterize joint structural, completion, and condition-binding challenges across input variations.

2. Related Work

2.1. Controllable Video Generation

Video generation has increasingly adopted spatial and temporal conditions to complement text and reference images. VideoComposer [27] combines multiple control signals, Motion Prompting [10] uses trajectories to guide scene motion, and SymphoMotion [32] jointly controls camera and object motion. VideoDirectorGPT [18] uses LLM-guided planning to maintain layout and entity consistency across multiple scenes. 3D Scene Prompting [17] uses scene geometry to maintain consistency across viewpoints. More directly related, Proxy2World [29] explores synthesizing detailed videos from lightweight structural proxies and the balance between structural adherence and visual quality. These methods demonstrate the value of spatiotemporal guidance. Our ProxyGenBench instead studies structural inputs that leave content unspecified or contain artifacts requiring correction.

2.2. Video Generation Benchmarks

Existing benchmarks assess complementary aspects of generated videos. VBench [14] and EvalCrafter [19] evaluate visual quality, temporal consistency, and semantic alignment; VBench-2.0 [33] and VideoPhy [5] extend evaluation toward physical realism and commonsense. VideoScore [13] and EvalVerse [30] further assess human-aligned and cinematic quality. T2V-CompBench [24] examines compositional requirements such as attribute binding and object interactions, while TC-Bench [9] focuses on temporal state and relation changes. Concurrent work PROWBench [15] evaluates fidelity to program-specified world states and events, including proxy-conditioned generation. ProxyGenBench instead focuses on imperfect structural proxies, requiring models to preserve intended spatial and temporal relationships, correct proxy artifacts, and bind supplementary conditions within the same scene. Varying proxy detail and complementary inputs lets us examine how these requirements interact.

3. ProxyGenBench

ProxyGenBench evaluates how different input conditions affect video generation. It tests whether models can preserve the intended scene structure, complete missing content and correct proxy artifacts, and integrate additional text, image, and audio conditions. These challenges often arise together, but a model may perform well on some while struggling with others.

3.1. Problem Setting and Structural Proxies

Structural proxy. A structural proxy is a video that specifies the intended scene layout, object placement, motion, camera behavior, and event order without fully defining the final result. Our proxies are rendered from 3D representations obtained from various sources. They may contain placeholder geometry, incomplete interactions, or unnatural motion. We use *proxy fidelity* to describe the level of geometric and motion detail, rather than content accuracy. A higher-fidelity proxy is thus not necessarily more accurate.

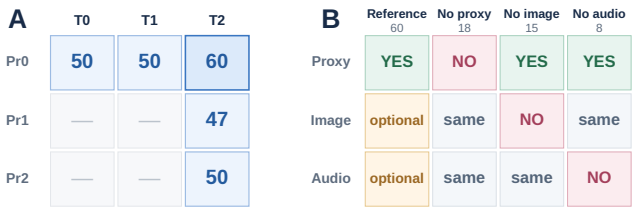


Figure 2. Evaluation case construction. (a) The proxy–text grid contains 257 cases; dashes mark omitted configurations. (b) Ablations add 18 no-proxy, 15 no-image, and 8 no-audio cases; the 60 reference cases are included in (a). The 14 targeted cases complete the benchmark: 312 cases overall, including 294 with proxy videos.

Generation setting. We define each evaluation case as $c = (s, p, U)$, where s is the intended scene specification, p is a structural proxy, and U includes a text prompt and optional image or audio references. Given (p, U), the model generates a video v. The scene specification s is used only for benchmark evaluation and is not provided to the video model. We derive evaluation checklists only from the conditions available to the model, ensuring that it is not evaluated on information it never received. Details are in the appendix.

3.2. Benchmark Construction

We construct the benchmark by varying several aspects of proxy-guided video generation. Scene tasks define the spatial and temporal relationships to preserve. Proxy sources and fidelity levels determine how much scene structure is provided and what remains incomplete. Additional text, image, and audio inputs introduce further requirements for the same scene. These variations test different aspects of the same generation problem rather than separate tasks.

Scene tasks. Dynamic action tests motion, interactions, and state transitions; multi-shot storytelling tests event order and continuity across cuts; and multi-view consistency tests stable scene and object states across viewpoints. These tasks cover motion, cross-shot continuity, and cross-view consistency; each task is also evaluated on the completion and reference-grounding requirements that apply to it. We additionally annotate five content categories—Characters, Navigation, Objects, Combat & Transform, and Camera—to describe coverage rather than create separate tasks.

Proxy sources. We collect proxies from four source groups, each with a different construction workflow (Fig. 3). Enhanced R2V adds proxy guidance to existing image- or audio-conditioned tasks. Native Code-to-Proxy uses an

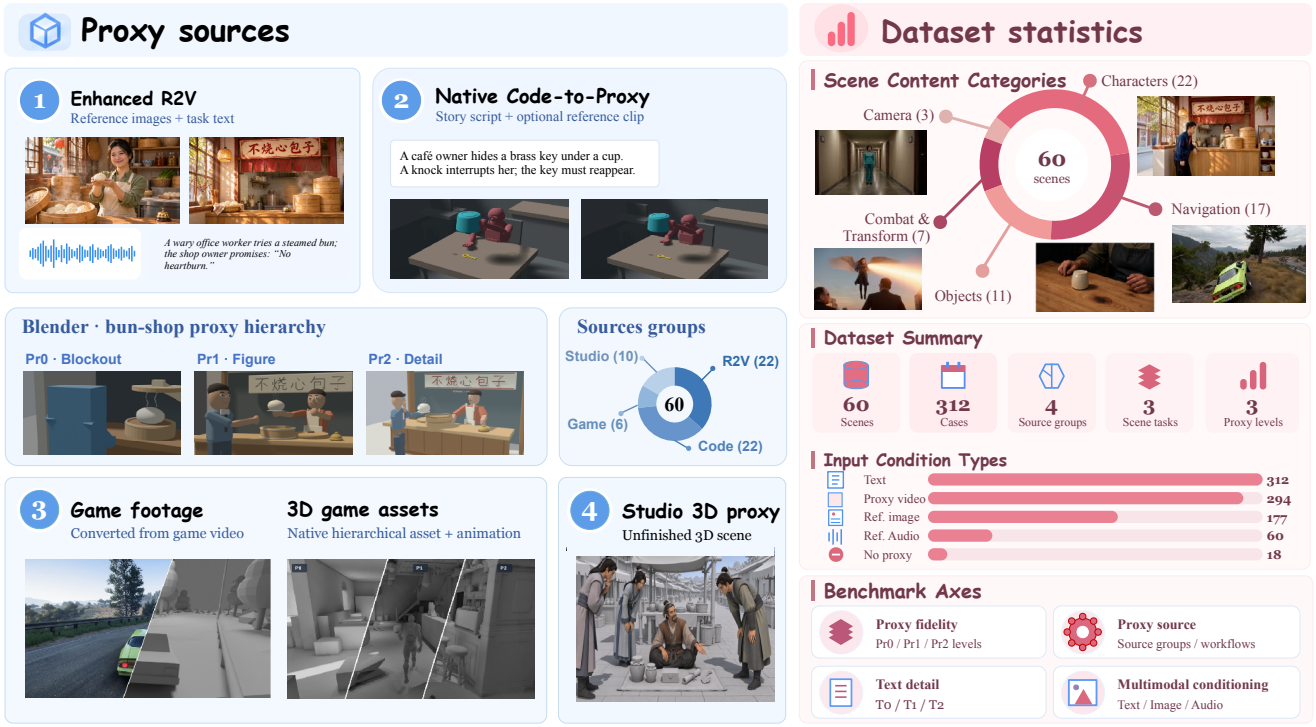


Figure 3. **Proxy sources and benchmark coverage**: 60 scenes and 312 cases from four source groups. Game Proxy includes game footage and 3D game assets. Source-group and content-category counts are scene-level; input-condition counts are case-level and non-exclusive.

LLM-based agent to animate 3D scenes from scripts or reference videos. Game Proxy includes game footage and 3D game assets, while Studio 3D Proxy uses unfinished production assets. Together, these sources cover a range of incomplete geometry and motion, with different types of artifacts across sources. We provide details of each construction workflow and its associated proxy artifacts in the appendix.

Input variations. We create multiple evaluation cases from each scene by varying the input conditions. We define three proxy-fidelity levels—Pr0 (Blockout), Pr1 (Figure), and Pr2 (Detail)—with increasing geometric and motion detail. Comparing these levels allows us to study whether more proxy detail improves both structural adherence and natural completion. Text-detail levels T0–T2 vary the amount of semantic guidance. Optional image and audio references add requirements for appearance, voice, and events. We vary text detail at a fixed proxy level (Pr0) and proxy fidelity at a fixed text level (T2) (Fig. 2). This allows us to study their effects separately without testing every combination.

Evaluation cases. We construct 312 cases from 60 scenes (Fig. 2), including 294 proxy-conditioned cases and 18 no-proxy counterparts. The proxy–text grid contains 257 cases, including 60 reference cases. We add 15 no-image, 8 no-audio, and 18 no-proxy cases, plus 14 targeted cases that vary joint reference removal, identity, voice, event timing, or camera control. We separately analyze the Studio 3D proxy subset across five models, while the targeted input variations are examined individually in the appendix. The configurations allow us to compare the effects of proxy availability, input detail, and reference grounding. Full case assignments and evaluation protocols are provided in the appendix.

3.3. Evaluation Framework

A proxy specifies only part of the intended video, so matching the proxy alone does not guarantee successful generation. Instead, we use case-specific checklists to assess whether the generated video retains the intended structure, completes or corrects unfinished content, and binds supplementary conditions to the correct entities and events (Table 1 and Fig. 4).

Evaluation dimensions. **Preserve** (P1–P3) tracks whether intended scene structure, motion, camera behavior, and event continuity survive generation. **Complete** (C1–C3) assesses whether the incomplete proxy is realized as coherent appearance, natural articulation, and plausible interactions instead of carrying over abstraction artifacts. **Bind** (B1–B2) assesses whether text, image, and audio conditions are correctly associated with entities, actions, and moments within that same structure. In the bakery example (Fig. 4), the model must preserve the seating arrangement and event order, turn rigid proxy gestures into natural interactions, and bind the specified clothing and voice references to the correct characters.

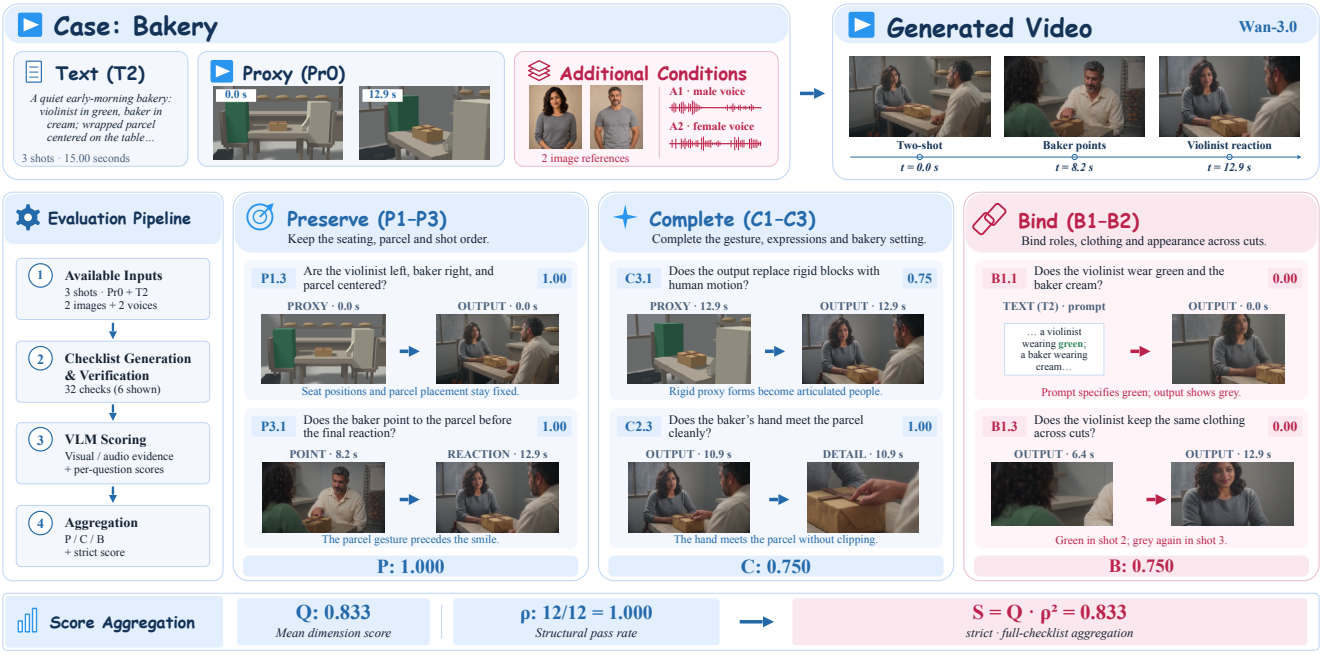


Figure 4. **Checklist-guided evaluation on the bakery case.** A case-specific checklist is generated and verified from the proxy, text, two image references, and two voice references, then used to score the Wan-3.0 video. Six of the 32 criterion-level assessments are illustrated with visual evidence, including the baker's parcel gesture and the violinist's clothing change across cuts. The full checklist yields $Q = 0.833$. All 12 critical Preserve criteria pass, giving $\rho = 12/12 = 1$ and Strict $S = Q\rho^2 = 0.833$.

Table 1. **Evaluation structure of ProxyGenBench.** Three dimensions comprise eight capabilities and 32 fine-grained criteria, with four criteria per capability. The table summarizes each capability's evaluation focus; criterion definitions are provided in the appendix. Case-specific checklist questions instantiate the applicable criteria.

Dimension	ID	Capability	Evaluation focus
Preserve	P1	Spatial Structure	Proxy-defined scene layout, relative object positions, and spatial anchors.
	P2	Motion & Camera Control	Proxy-defined coarse object trajectories, actions, and camera behavior.
	P3	Temporal & Shot Continuity	Proxy-defined event order and object-state continuity across transitions.
Complete	C1	Visual Realization Quality	Plausible appearance, geometry, texture, and visual coherence.
	C2	Motion & Physical Plausibility	Natural fine-grained articulation, physical interactions, and contact consistency.
	C3	Proxy Abstraction Handling	Completion of proxy placeholders and correction of proxy-induced artifacts.
Bind	B1	Entity & Appearance Grounding	Correct assignment of textual and visual attributes to entities.
	B2	Action & Audio-event Grounding	Correct assignment of actions, voices, and sound events to entities and moments.

The baker may point to the parcel at the right moment while the violinist still appears in the wrong outfit. This shows how preserving the intended structure can coexist with incomplete realization or incorrect condition binding.

Checklist construction. Eight capabilities comprise 32 fine-grained criteria: 12 for *Preserve*, 12 for *Complete*, and 8 for *Bind*. For each case, Gemini-3.1-Pro [11] instantiates applicable criteria into input-grounded questions [4, 25, 26], potentially with multiple subquestions about entities, events, or intervals. Qwen3.5-397B-A17B [22] verifies clarity, support from the available inputs, and alignment with the intended criterion; invalid questions are revised, with human feedback used to refine the procedure. Checklists are fixed before video scoring and shared across models on the same case. Each applicable criterion contributes one assessment.

Video scoring. The Gemini judge scores each criterion based on visual or audio evidence. Directly verifiable requirements receive binary scores, while appearance and motion quality are evaluated using predefined grading scales. In Fig. 4, the rigid-motion correction receives a graded score of 0.75 even though the event-order check passes, illustrating why proxy adherence alone does not determine completion quality. Reference-dependent criteria compare generated appearance, voices, or sound events with the supplied in-

Table 2. **Effect of proxy guidance.** (a) Control metrics on overlapping subsets (Pairs/Sc.: pairs/scenes; Δ: proxy minus no proxy; scores in $[0, 1]$). (b) VLM preferences over 50 pairs; Other includes ties and order-dependent outcomes.

(a) Spatiotemporal control

Metric	Pairs/Sc.	No proxy	+ Proxy	Δ
Action transition ↑	22/8	0.894	**1.000**	+0.106
Cross-shot spatial ↑	14/5	0.833	**0.952**	+0.119
MEt3R ↓	17/10	0.464	**0.337**	−0.127

(b) VLM preference

Protocol	Proxy win	Other	Proxy loss
Primary order	40	2	8
Two-order aggregate	**35**	15	0

puts. Inapplicable criteria are excluded; scoring prompts and rubrics are provided in the appendix.

3.4. Score Aggregation

Preserve, *Complete*, and *Bind* provide separate diagnostic scores. We further combine them into a structure-aware score, since even a visually realistic video may fail to meet the creator's intent if it breaks the intended scene layout or event order. We therefore introduce the *Strict* score, which places greater emphasis on critical *Preserve* requirements while retaining the three separate dimension scores.

Hierarchical aggregation. Each applicable criterion contributes one score, regardless of the number of subquestions. For case i, let $s_{i,k} \in [0, 1]$ denote the score for criterion k. Using equal weights $a_k = 1$, applicable criteria are aggregated into capability scores:

$$f_{i,g} = \frac{\sum_{k \in \mathcal{K}_{i,g}} a_k s_{i,k}}{\sum_{k \in \mathcal{K}_{i,g}} a_k}. \tag{1}$$

Here, $\mathcal{K}_{i,g}$ contains the criteria applicable to capability g; inapplicable criteria are omitted. The resulting capability scores form *Preserve*, *Complete*, and *Bind* dimension scores $D_{i,d}$ using the fixed aggregation specified in the appendix.

Strict Score and Joint Success Rate. We use the 12 predefined *Preserve* criteria (P1.1–P3.4) as critical structural checks, with pass thresholds set before evaluation. For each case i, let $\mathcal{D}_i$ be the set of applicable dimensions and ρ_i be the fraction of the 12 critical *Preserve* checks that pass. We define the *Average* score Q_i and *Strict* score S_i as follows:

$$Q_i = \frac{1}{|\mathcal{D}_i|} \sum_{d \in \mathcal{D}_i} D_{i,d}, \qquad S_i = Q_i \rho_i^2. \tag{2}$$

The factor ρ_i^2 gives greater weight to structural preservation: failing multiple critical checks reduces the overall score, even if the video looks complete or satisfies other conditions. Thus, *Strict* emphasizes proxy control while keeping *Preserve*, *Complete*, and *Bind* as separate diagnostic scores. We discuss this weighting and alternatives in the appendix. *Joint Success* rate measures the fraction of outputs that pass all applicable critical *Preserve* checks ($\rho_i = 1$), rather than all P/C/B requirements. We compute model-level *Strict* by averaging case scores within each input group. Missing model outputs receive a score of zero, while cases that cannot be evaluated are excluded.

Table 3. **Overall proxy-conditioned performance.** $P/C/B$ denote Preserve/Complete/Bind; Average Q is the mean of applicable dimensions and Strict $= Q \times (\text{critical pass rate})^2$. Joint is the fraction passing all 12 critical Preserve criteria. Missing model outputs receive zero. Scores lie in $[0, 1]$ (higher is better); bold marks the best score within each group.

Model	P	C	B	Average	Strict	Joint
(a) Without independent audio						
Seedance-2.5	**0.774**	0.566	**0.693**	**0.678**	**0.450**	0.140
Kling-4.0	0.746	**0.591**	0.652	0.663	0.425	0.150
Wan-3.0	0.732	0.557	0.642	0.644	0.399	0.164
Seedance-2.0	0.698	0.505	0.642	0.615	0.375	**0.174**
MiniMax-H3	0.639	0.539	0.667	0.615	0.344	0.072
(b) With independent audio						
Seedance-2.5	**0.645**	**0.612**	**0.740**	**0.666**	**0.325**	0.040
Seedance-2.0	0.633	0.550	0.718	0.634	0.314	**0.080**
Wan-3.0	0.617	0.511	0.624	0.584	0.324	**0.080**
MiniMax-H3	0.553	0.536	0.654	0.581	0.241	0.020

4. Experiments

We first assess the control benefits of structural proxies, then examine joint structural constraints, proxy fidelity, text detail, and condition binding. We finally compare the VLM evaluator with human judgments.

4.1. Experimental Setup

Models and protocol. We evaluate Seedance-2.5 [8], Kling-4.0 [16], Wan-3.0 [1], Seedance-2.0 [7], and MiniMax-H3 [20] on supported configurations among the 294 proxy-conditioned cases. Results are grouped by independent audio-reference availability; Kling-4.0 is excluded from the audio-reference group. Models within each group share cases and verified checklists. *Strict* provides the primary structure-sensitive ranking, while the 18 no-proxy cases are used for matched comparisons. Missing model outputs receive zero, as specified in Sec. 3.4. Model settings and evaluation details are provided in the appendix.

Proxy availability. We evaluate 50 model-level matched proxy/no-proxy video pairs from 18 scenes, holding text, other references, and generation settings fixed. Action-transition, spatial-consistency, and geometric metrics use

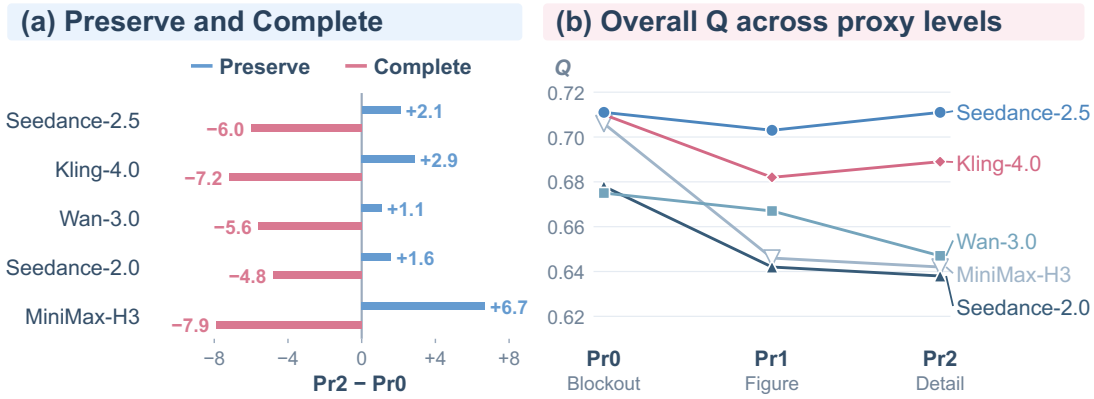


Figure 5. **Proxy-fidelity effects on shared T2 scenes.** (a) Preserve and Complete changes from Pr0 to Pr2, in percentage points. (b) Mean dimension score Q across proxy levels. The full $P/C/B/Q$ results appear in Table 4.

Table 4. **Results across three proxy-fidelity levels.** $P/C/B$ denote Preserve/Complete/Bind, and **Average** is their mean. For each model and metric, we **bold** the best score across fidelity levels to highlight which level performs best for each metric.

	Pr0 – Blockout				Pr1 – Figure				Pr2 – Detail			
Model	P	C	B	Average	P	C	B	Average	P	C	B	Average
Seedance-2.5	0.780	**0.598**	0.754	**0.711**	**0.801**	0.597	0.714	0.704	**0.801**	0.538	**0.794**	**0.711**
Kling-4.0	0.753	**0.647**	**0.730**	**0.710**	0.774	0.610	0.661	0.682	**0.782**	0.575	0.710	0.689
Wan-3.0	0.755	**0.587**	**0.681**	**0.675**	**0.782**	0.553	0.665	0.667	0.766	0.531	0.645	0.647
Seedance-2.0	0.737	**0.560**	**0.738**	**0.678**	**0.788**	0.482	0.657	0.642	0.753	0.512	0.649	0.638
MiniMax-H3	0.664	**0.612**	**0.843**	**0.706**	**0.737**	0.528	0.673	0.646	0.731	0.533	0.661	0.642

their applicable, potentially overlapping subsets. VLM preferences are aggregated across both presentation orders to reduce order bias.

Cross-condition comparisons. Adding text detail or references can change the requirements applicable to a case, especially for Bind. We therefore compare matched configurations on shared checklist criteria and assess newly introduced binding requirements separately. Scores from different case sets or checklists are not interpreted as isolated modality effects. Details are in the appendix.

4.2. Proxy Benefits

The structural prior helps when prompts leave temporal and spatial relationships ambiguous (Table 2). On matched subsets, action-transition accuracy rises from 0.894 to 1.000, cross-shot spatial consistency from 0.833 to 0.952, and MEt3R [3] error falls from 0.464 to 0.337. Under the two-order VLM protocol, proxy-guided outputs are preferred in 35 of 50 pairs (70%); the other outcomes are ties or order-dependent. These comparisons establish benefits for the measured structural controls, while motion quality and condition binding require separate evaluation.

4.3. Joint Structural Success

Average *Preserve* scores do not show whether all critical structural checks hold in the same output (Table 3). Without independent audio references, Seedance-2.5 and Kling-4.0 lead *Strict* (0.450 and 0.425), yet their *Joint Success* rates are only 14.0% and 15.0%. Their strengths also differ: Seedance-2.5 leads Preserve and Bind, whereas Kling-4.0 leads Complete. Across all five models, *Joint Success* ranges from 7.2% to 17.4%, with Seedance-2.0 highest despite a lower *Strict* score. This ranking reversal reveals a gap between structure-sensitive overall performance and joint constraint satisfaction: a model can achieve a higher *Strict* score without satisfying all critical structural constraints together.

4.4. Proxy Fidelity

More detailed proxies do not consistently improve the final video (Table 4 and Fig. 5). On the 31 shared T2 scenes without independent audio references, *Preserve* increases by 1.1–6.7 percentage points from Pr0 to Pr2 across all five models, while *Complete* decreases by 4.8–7.9 points. This tension persists even in the leading models: Seedance-2.5 gains 2.1 points in *Preserve* but loses 6.0 in *Complete*, while Kling-4.0 gains 2.9 but loses 7.2. Their overall responses differ, however. Seedance-2.5 maintains an Average score of 0.711 as *Bind* improves from 0.754 to 0.794, whereas Kling-4.0 drops from 0.710 to 0.689. Pr0 yields the highest *Average* for the other four models. Thus, stronger overall performance does not eliminate the preservation–completion tension, but models differ in how they balance these requirements.

More detailed proxies specify useful structure but may

Table 5. **Strict scores across input configurations.** (a) Text-detail levels at Pr0 without independent audio references; bold indicates the highest score for each model. (b) Multimodal configurations, where Pr, T, I, and A denote proxy, text, image references, and independent audio references. The +I, +A, and +I+A columns add the indicated references to Pr+T. '-' denotes an unsupported input.

(a) Text detail				(b) Multimodal inputs				
Model	**T0**	**T1**	**T2**	**Model**	**Pr+T**	**+I**	**+A**	**+I+A**
Seedance-2.5	0.434	0.454	**0.458**	Seedance-2.5	**0.454**	0.391	0.365	0.256
Kling-4.0	0.394	0.426	**0.437**	Kling-4.0	**0.449**	0.347	–	–
Wan-3.0	0.348	0.382	**0.425**	Wan-3.0	**0.438**	0.352	0.349	0.394
Seedance-2.0	0.325	**0.385**	0.375	Seedance-2.0	**0.431**	0.265	0.199	0.191
MiniMax-H3	0.307	0.336	**0.375**	MiniMax-H3	**0.392**	0.290	0.252	0.266

Table 6. **Performance on Studio 3D proxy cases.** P, C, and B denote Preserve, Complete, and Bind; Strict $= Q \times$ critical pass rate^2. Scores lie in $[0, 1]$ (higher is better); bold marks the highest score in each column.

Model	P	C	B	Strict
Seedance-2.5	0.924	0.511	0.318	0.505
Kling-4.0	0.894	0.633	0.318	0.525
Wan-3.0	0.947	**0.665**	0.273	**0.564**
Seedance-2.0	**0.962**	0.477	0.182	0.498
MiniMax-H3	0.848	0.616	**0.500**	0.536

Table 7. **Human validation on 3,199 judgments from 110 videos.** Judg. counts audited judgments; Valid and Acc. denote checklist validity and scoring accuracy against human annotations. Pass and Fail report accuracy within judge-labeled pass and fail groups.

Dimension	**Judg.**	**Valid**	**Acc.**	**Pass**	**Fail**
Preserve	1,320	0.998	0.897	0.902	0.865
Complete	1,320	1.000	0.952	0.963	0.916
Bind	559	1.000	0.948	0.981	0.840
Overall	**3,199**	**0.999**	**0.929**	**0.939**	**0.885**

also contain rigid poses, awkward contacts, or unnatural motion. Following those local details too closely can carry proxy artifacts into the final video instead of correcting them. These opposing trends highlight a central challenge: generators must follow scene-defining constraints without copying unintended proxy details.

4.5. Text Detail

Text supplies semantic information that proxy geometry and motion leave unspecified (Table 5(a)). Under Pr0 without independent audio references, T2 scores higher than T0 on *Strict* for all five models, although Seedance-2.0 peaks at T1. The observed T0-to-T2 gains are larger for Wan-3.0 (+7.7 points) and MiniMax-H3 (+6.8) than for Seedance-2.5 (+2.4), narrowing their gaps to the leading model. These gains suggest that detailed text can compensate for weaker baselines, rather than benefiting all models equally.

4.6. Condition Binding

Following the proxy does not guarantee that supplementary conditions are assigned to the intended entities and events.

Image references. On their applicable subsets, identity matching and cross-shot identity consistency average 0.897 and 0.909, whereas action–object binding averages only 0.636. Some generated videos swap face identities while retaining the original clothing assignments. Thus, recognizing reference attributes is distinct from binding those attributes to the right entities and actions. All five models also obtain lower observed Strict scores on Pr+T+I cases than on Pr+T cases (Table 5(b)), reinforcing the difficulty of integrating visual references into an already structured scene (see the appendix for detailed analyses).

Audio references. Voice/source matching averages 0.768, while audiovisual timing alignment reaches only 0.513. A model may reproduce the intended voice without placing it at the correct moment. This reveals a temporal binding failure that structural guidance alone does not resolve.

Studio 3D proxies. On the 12 Studio 3D proxy cases evaluated across all five models (Table 6), *Preserve* ranges from 0.848 to 0.962, whereas *Bind* reaches only 0.182–0.500. Seedance-2.5 and Kling-4.0, the overall Strict leaders without independent audio, both score $B = 0.318$, below MiniMax-H3 ($B = 0.500$). Seedance-2.0 attains $P = 0.962$ but only $B = 0.182$. Proxy fidelity reveals a similar separation: *Preserve* rises from Pr0 to Pr2, yet *Bind* drops by 3.6 points for Wan-3.0 and 18.2 for MiniMax-H3 (Table 4). Together, these results show that condition binding is not a by-product of structural control: models must ground supplementary information within the structure they preserve.

4.7. Human Validation

We audit 3,199 criterion-level judgments from 110 videos (Table 7). Checklist validity is 99.9%, and overall scoring accuracy against human annotations reaches 92.9%. Scoring

accuracy is lowest for Preserve (89.7%); for Bind, accuracy on judge-labeled failures is 84.0%, compared with 98.1% for passes. This gap matters because the benchmark must detect subtle structural violations and incorrect grounding. We therefore report dimension-level and pass/fail accuracy rather than relying only on an aggregate score; further annotation details are provided in the appendix.

5. Conclusion

ProxyGenBench evaluates three interacting requirements of video generation from imperfect structural proxies: Preserve, Complete, and Bind. Across five models, proxy guidance improves spatiotemporal control, yet strong average adherence can mask violations of joint structural constraints. More detailed proxies can also reduce completion scores, while preserved structure does not ensure correct condition binding. These findings highlight the need for selective control: generators must retain intended structural constraints while correcting proxy artifacts and grounding supplementary information in the same scene.

References

[1] Alibaba Cloud. Wan 3.0. https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model, 2026. 6

[2] Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16173–16184, 2026. 2

[3] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2025. 7

[4] Xuehai Bai, Zhenchen Tang, Yang Shi, Dianyi Wang, Tengfei Liu, Wanshun Su, Xuanyu Zhu, Ruohui Wang, Haiwen Diao, Haotian Wang, et al. Think before you score: Thinking reward model for visual generation. *arXiv preprint arXiv:2609.37372*, 2026. 5

[5] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, pages 102075–102121, 2025. 3

[6] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 conference papers*, pages 1–11, 2024. 2

[7] ByteDance Seed Team. Seedance 2.0. https://seed.bytedance.com/en/blog/official-launch-of-seedance-2-0, 2026. 6

[8] ByteDance Seed Team. Seedance 2.5. https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, 2026. 6

[9] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024. 3

[10] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 3

[11] Google DeepMind. Gemini 3.1 Pro model card, 2026. 5

[12] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, pages 100433–100464, 2025. 2

[13] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024. 3

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[15] Zheng-Hui Huang, Guixu Lin, Yu-Ju Tsai, Jian-Kai Zhu, Fengbo Lan, Yu-Lun Liu, Yung-Yu Chuang, Kaipeng Zhang, and Zhixiang Wang. Prowbench: Do video models render what the program specifies?, 2026. 3

[16] Kuaishou Technology. Kling 4.0. https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-officially-launches-next-generation-kling-40-raising, 2026. 6

[17] JoungBin Lee, Jaewoo Jung, Jisang Han, Takuya Narihira, Kazumi Fukuda, Junyoung Seo, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. 3d scene prompting for scene-consistent camera-controllable video generation, 2025. 3

[18] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3

[19] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149. IEEE, 2024. 3

[20] MiniMax-AI. Minimax-h3. https://github.com/MiniMax-AI/MiniMax-H3, 2026. 6

[21] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[22] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. 5

[23] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 2

[24] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8406–8416. IEEE, 2025. 3

[25] Zhenchen Tang, Yang Li, Songlin Yang, Bo Peng, Xiaotong Zhao, Shuai Li, Haotian Fan, Alan Zhao, and Jing Dong. Rewardverse: Rubric-guided policy optimization for video reward modeling. *arXiv preprint arXiv:2609.22947*, 2026. 5

[26] Zhenchen Tang, Bo Peng, Zichuan Wang, Songlin Yang, Leilei Cao, Fengjie Zhu, and Jing Dong. An evolutionary agentic approach for open-ended image quality perception. *arXiv preprint arXiv:2609.22942*, 2026. 5

[27] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems*, pages 7594–7611. Curran Associates, Inc., 2023. 3

[28] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[29] Hongli Xu, Weilong Yan, Anbang Wang, Chunyu Zou, Siyu Hong, and Jingwei Huang. Proxy2world: Learning to generate worlds from lightweight proxies without seeing them, 2026. 3

[30] Songlin Yang, Haobin Zhong, Ruilin Zhang, Xiaotong Zhao, Shuai Li, Kai Zheng, Xuyi Yang, Zhe Wang, Zhenchen Tang, Yang Li, et al. Evalverse: Pipeline-aware and expert-calibrated benchmarking for professional cinematic video generation. *arXiv preprint arXiv:2605.23271*, 2026. 3

[31] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, pages 83048–83077, 2025. 2

[32] Guiyu Zhang, Yabo Chen, Xunzhi Xiang, Junchao Huang, Zhongyu Wang, and Li Jiang. Symphomotion: Joint control of camera motion and object dynamics for coherent video generation. *arXiv preprint arXiv:2604.03723*, 2026. 3

[33] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. 3

ProxyGenBench: Benchmarking Proxy-Guided Video Generation

Supplementary Material

A. Benchmark Construction

A.1. Proxy and Text Fidelity

Proxy levels. For each scene, we define up to three proxy levels, Pr_0, Pr_1, and Pr_2. When multiple levels are available, they preserve the same story, shot structure, camera plan, blocking, role colours, and key trajectories, while varying primarily in geometric detail and motion fidelity. For Blender-authored proxies, Pr_0 uses coarse coloured primitives; Pr_1 adds articulated bodies and readable props; and Pr_2 remains untextured but provides clearer geometry and more faithful motion. For proxies from other sources, the available levels may have source-specific representations and may not span all three levels; nevertheless, they follow the same abstract-to-detailed ordering, with Pr_0 denoting the most abstract representation, as shown in Fig. 6.

Text prompt levels. We define three text-prompt levels, T_0, T_1, and T_2, which progressively increase the amount of semantic and temporal information supplied beyond the proxy video (Figure 7). T_0 provides a concise target description—including the intended visual domain, principal subjects, setting, and key transformation—while relying primarily on the proxy for shot order, camera framing, blocking, and motion. T_1 additionally specifies the desired appearance of characters, objects, and environment, together with coarse shot-level actions and temporal beats. T_2 gives the most complete specification: detailed visual attributes, physical and interaction constraints, cinematographic instructions, negative constraints, audio requirements when applicable, and a time-resolved plan aligned with the proxy. Across all levels, the proxy remains the primary structural signal; the prompt directs the model to recreate the intended scene with the desired appearance and realism.

Controlled case construction. Each scene is organized as a controlled experiment with a shared target story, duration, generation settings, and reference conditions. The possible core configurations (when the corresponding proxy levels are available) are

$$\{Pr_0\text{-}T_0,\ Pr_0\text{-}T_1,\ Pr_0\text{-}T_2,\ Pr_1\text{-}T_2,\ Pr_2\text{-}T_2\}.$$

The Pr_0-$T_{0,1,2}$ cases vary text specificity at fixed proxy fidelity, while $Pr_{0,1,2}$-T_2 varies proxy detail and motion fidelity at fixed detailed text.

In addition to the proxy–text grid and the input ablations described below, we include 14 targeted control cases. Each keeps the same scene goal but changes one specific input or instruction. Four cases test voice assignment in two dialogue scenes: in the normal version, each character uses their designated voice reference; in the swapped version, the two characters exchange voice references. The remaining cases test whether models follow a specific identity reference, assign the correct image to the correct role, delay the closing of a gate, distinguish camera movement from zoom, shift focus at the intended time, keep a centered camera while moving through a doorway, or operate without both image and audio references. These controls test whether a model uses each condition for its intended purpose rather than relying on an easier but incorrect cue.

A.2. Proxy Construction

Source workflows. Proxies originate from four source groups. *Enhanced R2V* adds animated structural guidance to reference-to-video tasks; *Native Code-to-Proxy* animates three-dimensional scenes from scripts or optional reference clips through an LLM-based authoring process; *Game Proxy* uses game footage or native game assets and animation; and *Studio 3D Proxy* uses unfinished production scenes. These groups provide heterogeneous incompleteness and artifact patterns, rather than interchangeable samples of one authoring pipeline. The four workflows span reference-conditioned content, script-authored scenes, game environments, and unfinished animation assets, providing distinct sources of missing detail and imperfect motion.

Blender-authored proxies and native hierarchies. In the Blender-authored examples, lower-detail proxies contain color-coded primitives, while more detailed versions replace these with articulated figures, distinguishable props, and more informative local motion. Native 3D hierarchies and unfinished studio assets retain source-specific geometric representations and animation. When multiple levels are present, the target story and major camera and event structure are kept consistent; not every source provides every fidelity level. The level labels describe the amount of structural and motion detail, not the correctness of all depicted details.

A.3. Scene Coverage

Scene tasks identify the principal structural challenge: dynamic action, multi-shot storytelling, or multi-view consistency. Separately, five content-category annotations describe scene subject matter: Characters (22), Navigation (17), Objects (11), Combat & Transform (7), and Camera (3). The proxy-source, task, and content-category annotations serve distinct roles and should not be interpreted as three mutually exclusive partitions of the same type.

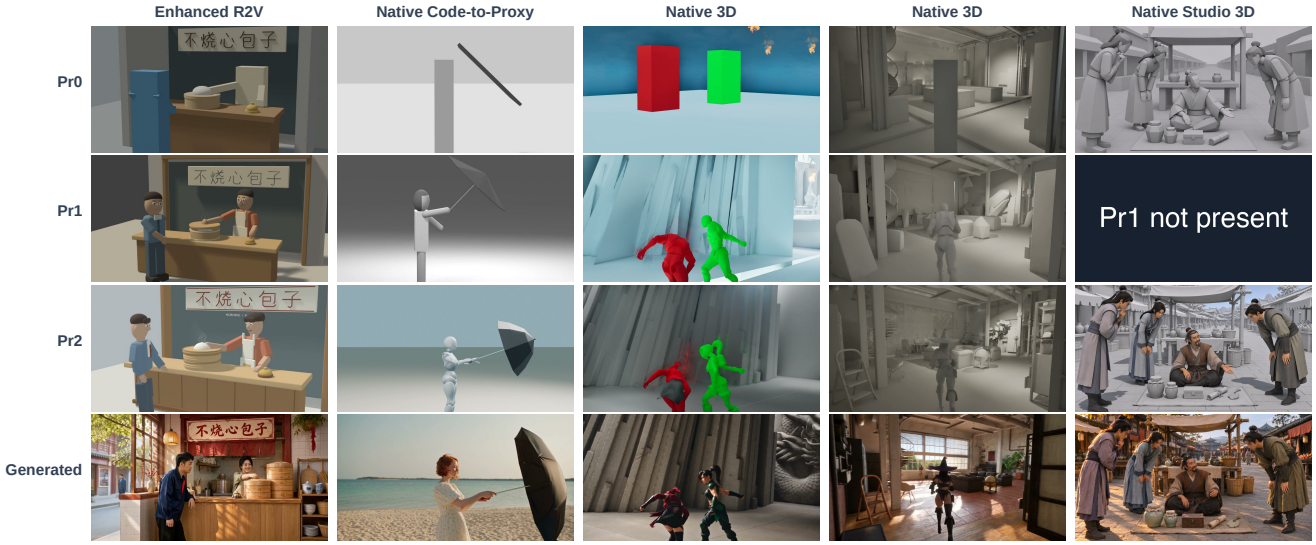


Figure 6. Representative proxy fidelity levels and corresponding Seedance-2.5 outputs. The Studio 3D example has no Pr_1 proxy.

Table 8. Benchmark configurations and controlled comparisons. Text and proxy level counts follow the order $T_0/T_1/T_2$ and $Pr_0/Pr_1/Pr_2$, respectively. The fidelity analysis uses 31 matched scenes.

Contrast	**Cases**	**Fixed conditions**	**Research question**
Text detail	Pr_0: 50/50/60	Proxy fidelity, scene intention	Does more explicit text improve realization of the proxy-defined sequence?
Proxy detail	T_2: 60/47/50	Target text, scene, references	How does added geometric and motion detail affect preservation and completion?
Reference removal	15 no-image; 8 no-audio	Scene and remaining conditions	Which identity, appearance, voice, and event requirements depend on references?
Proxy removal	18 no-proxy	Target description, remaining references, generation settings	How much does structural guidance help motion, camera, and continuity?
Targeted controls	14 variations	Base scene apart from the targeted change	Are identity, voice, event timing, and camera instructions grounded correctly?

The experimental contrasts are organized in Table 8. The 257 core cases are augmented with 18 no-proxy, 15 no-image, eight no-audio, and 14 targeted cases, giving 312 cases from 60 scenes. Of these, 294 include a proxy video and 18 omit it. Image and independent audio references appear in 177 and 60 cases, respectively; reference-type counts are non-exclusive. The Studio 3D subset contains 12 cases from 10 scenes and is analyzed separately in the main paper; these cases are part of the 312-case benchmark, rather than additional targeted variations. The case identifier, source, task, proxy and text levels, and optional reference indicators define the corresponding experimental configuration.

A.4. Target Conditions

Text prompts communicate intended appearance, interactions, sound, and target style in addition to any structural requirements conveyed by the proxy. Reference images can specify identity, clothing, objects, and appearance, while independent audio references can specify voice/source identity or sound events. Such requirements are evaluated for the entity or event to which they are assigned, not merely for their appearance somewhere in the video.

Intended stylization versus artifacts. Proxy fidelity is not a style target by default. Intended stylization, including intentionally simplified geometry or stylized movement, is described in the accompanying text and is not counted as an unintended proxy artifact. By contrast, proxy placeholders, rigid pseudo-actions, clipping, or missing articulation are completion or correction targets only when the generator's supplied conditions support that expectation. A checklist must not introduce stylistic requirements from the hidden scene specification alone.

Condition roles. The T0–T2 umbrella progression in Fig. 7 is an actual illustrative input from the benchmark. It exposes a useful separation between *where and when* (primarily controlled by the proxy), *what must be realized or repaired* (the target prompt), and *which entity receives a ref-*

Representative prompt-level progression. All examples below use the same low-detail proxy Pr_0, the same umbrella reference image, and the same 15-second target. They illustrate how increasing text specificity complements—rather than replaces—the proxy's shot and motion signal.

T_0 — concise scene directive

Use the attached proxy video for shot order, timing, camera framing and rough body and umbrella motion; use reference image 1 for the canopy's black colour, eight-rib geometry, curvature and glossy finish. Replace the proxy figures and background with natural live action.
Create a 15-second warm live-action beach advertisement. One young fair-skinned woman with short wavy auburn hair wears a white knee-length dress with small yellow lemon-floral prints and puff sleeves. She has one folding umbrella with a glossy black eight-rib canopy, silver-grey shaft and matte black handle. On a pale sandy beach beside a calm turquoise sea, she turns toward the camera, opens the umbrella, moves with it and brings it above and behind her. Follow the proxy for the changing views and gestures, rendering the umbrella mechanism and her movements naturally. Keep her appearance and the umbrella consistent. Include soft breeze, distant waves and an opening pop; no music.

T_1 — controlled appearance and coarse timeline

Treat the attached proxy reference video as guidance for motion, timing, framing and camera placement, not final appearance. Preserve its four beats, cut order, body orientations and umbrella positions without tracing rigid block movements. Reference image 1 supplies the canopy's glossy black colour, eight-rib geometry, curvature and finish only; exclude its white studio background.
Generate a 1280 × 720, 15-second polished, warmly graded live-action beach advertisement. Exactly one young woman stands alone with one folding umbrella. Keep her identity, costume and the umbrella's construction consistent across all cuts. Use pale beige sand, a calm blue-green sea, a thin distant coastline and a clear lightly hazed sky. Movement is subtle and human; sound is diegetic sea breeze, distant waves and an umbrella-opening pop; no music.
0.00–about 3.67: medium view; she turns gradually from a rear-facing pose toward camera, with the umbrella out of view. **About 3.67–7.50:** side view; she presents the folded umbrella, opens it, then raises it overhead. **About 7.50–11.25:** wider nearly frontal view; she moves the open umbrella beside her and lifts it outward and upward. **About 11.25–15.00:** closer view; she turns toward camera with the open umbrella behind her shoulder and settles into a front-facing hero composition.

T_2 — full visual, physical, and temporal specification

Use the attached proxy reference video only for motion, camera placement and timing. Preserve all four beats, cut order, framing, angles, body orientations and umbrella progression: out of view at first, then presented and opened, moved outward, and brought behind her shoulder. The proxy does not define appearance: replace its cuboid, slab, rigid motion, flat shading and two-colour backdrop with natural live action.
Create a 1280 × 720, 25-fps, 15-second commercial-grade lifestyle beach advertisement. The woman has fair skin, short softly wavy auburn hair, a peaceful expression, and a white summer dress with yellow lemon-floral prints and short puff sleeves. Require natural skin, credible hands and limbs, quiet breathing, small weight shifts, and breeze-responsive hair and fabric. The folding umbrella begins compact and closed; its working slider and articulated, exactly eight ribs open once, and its glossy black fabric must retain curved, tensioned panels rather than becoming metallic or grey. Use warm natural light, soft coherent shadows, atmospheric depth, pale grainy sand, a turquoise blue-green sea, and a low distant coastline. Keep the woman alone; no unrelated activity, weather change, or music.
0.00–about 3.67: medium view, rear-facing turn toward camera; umbrella absent. **About 3.67–7.50:** side view, plausible slider-and-rib opening with natural hand contact and one soft pop. **About 7.50–11.25:** wider nearly frontal view, the open umbrella travels beside her then outward and upward as her free arm extends. **About 11.25–15.00:** closer rear-three-quarter-to-front turn, ending in a relaxed hero pose; canopy, eight ribs, face, and hands remain readable.

Figure 7. Text-detail levels for the UMBRELLA GIRL case. The proxy and reference image remain fixed across T_0, T_1, and T_2.

erence (supplementary conditions). Table 9 records these distinct sources of evaluation evidence. In particular, a low-poly look is not an error when stylization is explicitly requested; a rigid placeholder is a completion target when natural articulation is specified. This distinction is established before viewing model outputs.

Audio references. Audio-related checks distinguish source assignment (who speaks or produces a sound), event correctness (whether the specified sound occurs), and synchronization (whether the sound aligns with the intended visible event). Independent audio-reference configurations are reported separately from configurations without independent

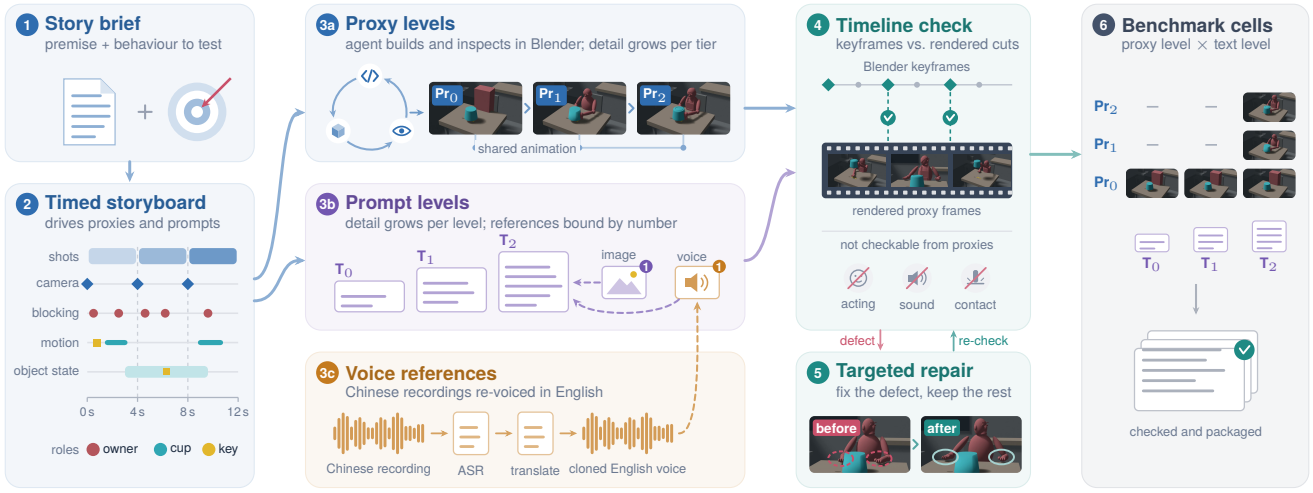


Figure 8. Proxy construction and repair pipeline, illustrated with *The key under the cup*, a case without image or audio conditions. Optional reference branches show steps used for other applicable scenes. Checks and repairs are performed iteratively, with scene-specific evidence.

Table 9. Input conditions and evaluation roles.

Input	Intended constraint	Evaluation
Proxy	Layout, motion, camera, event order	Preserve
Text	Appearance, style, actions, corrections	Complete; Bind
Image	Identity and appearance of named targets	Bind (B1)
Audio	Voice/source, sound event, timing	Bind (B2)

audio references, because input support and the applicable Bind requirements differ. Kling-4.0 is excluded from the audio-reference group because that input configuration is unsupported.

A.5. Agent-Assisted Construction and Verification

Our construction process combines LLM-assisted authoring with iterative inspection and correction, as illustrated in Fig. 8. An agent converts a shared scene specification into executable Blender scenes and English prompts. We then inspect the rendered proxies, source files, and packaged inputs, correcting inconsistencies before finalizing the benchmark cases. The figure illustrates the general workflow, although the inspection process and available evidence vary across scenes. Native 3D and Studio 3D sources use separate construction pipelines.

Shared scene specification. Construction begins with a story brief and the behavior to be tested, such as object persistence, role correspondence, or camera-motion control. The agent expands the brief into a timed storyboard specifying shot order, camera framing and movement, character blocking, key trajectories, and relevant object-state changes. This storyboard guides both proxy construction and prompt authoring (Fig. 8, steps 1–2). Stable role colours identify characters and salient objects within the proxies, while text and appearance references specify their intended final appearance.

Proxy-level construction. For Blender-authored scenes, the agent receives explicit instructions for each available fidelity level (Fig. 8, step 3a). At Pr_0, characters use coarse body blocks and root motion, with minimal arms or contact markers when essential to communicate an interaction. At Pr_1, articulated heads, torsos, and limbs provide clearer poses and object interactions. At Pr_2, additional geometry improves hands, feet, prop silhouettes, and local scene readability, together with fuller motion. Materials remain simple and untextured across these levels.

The execution harness runs the agent's generated `bpy` code in Blender and returns execution output, errors, scene information, and requested preview renders. The agent uses this feedback to construct and repair the scene incrementally before the final animation render. Where geometry-swap mode is used, a new tier starts from an existing animated proxy: replacement geometry inherits the original object's animation through a shared parent, while instructions preserve cameras, frame range, frame rate, timeline structure, role colours, and existing trajectories. These constraints guide the intended differences between levels; rendered inspection remains necessary to detect unintended changes. Native game and studio assets retain their available source representations and do not necessarily follow this Blender-authoring procedure.

Text-level construction and reference binding. The agent generates three levels of English prompts from the

shared storyboard (Fig. 8, step 3b). Each proxy-conditioned prompt begins with instructions on how to use the reference video, followed by a description of the intended final video. T_0 provides a broad semantic and visual goal, relying mainly on the proxy for spatial and temporal structure. T_1 adds appearance details and coarse descriptions of actions and shots. T_2 provides detailed requirements for appearance, interactions, camera work, timing, and sound when applicable. We review the prompt content to maintain these distinctions and shorten overly detailed T_0 prompts when necessary.

The proxy guides scene layout, camera motion, timing, and object movement, while text and appearance references specify characters, materials, environments, and visual style. Image and audio references are numbered according to their order in the input arrays. Input ablations retain the same target scene and remaining references, with prompts specifying the conditions available in each case. Appearance and voice requirements are expressed in text where needed in the absence of the corresponding reference media.

We use increasing character budgets as authoring guidelines: approximately 1,000–2,000 for T_0, 2,500–3,500 for T_1, and 4,000–5,000 for T_2, including proxy and reference instructions. Prompt lengths reflect scene complexity and the information required at each level. The same finalized case prompt is used across supported models.

Optional voice-reference conversion. For some existing R2V cases with Chinese voice references, we automatically transcribe and translate the recordings into English, then synthesize English speech using voices cloned from the original recordings (Fig. 8, step 3c). These translated recordings serve as voice references, not as dialogue that the generated video must reproduce.

The conversion targets speaker identity, intended meaning, and approximate speaking duration rather than exact word-level timing. Processed recordings are assigned to the corresponding numbered audio conditions and serve as references for generation and evaluation. We retain original recordings and conversion metadata where available and leave mixed-speaker or speech-with-music recordings unchanged.

Timeline verification with Blender sources. For audited scenes with available Blender files, we check explicit prompt timelines against both the source scene and the rendered proxy (Fig. 8, step 4). Source inspection includes frame range and frame rate, camera assignments, timeline markers, relevant animation keyframes, and, where needed, sampled camera transforms and focal lengths. These provide supporting evidence for intended event timing, shot boundaries, and camera behavior. Frame indices are converted to seconds using the scene's starting frame and frame rate.

We cross-check source evidence with sampled video frames, including adjacent frames around proposed cuts and frames surrounding important actions or state changes. A timeline marker may describe an action rather than a cut, and a camera assignment need not produce a visible cut when successive views coincide. We therefore distinguish source annotations from events visible in the rendered video. When the evidence supports a continuous shot, the prompt describes successive action beats without introducing additional cuts.

These checks establish structure only to the extent supported by the proxy. Coarse geometry may show a trajectory or an object's broad location without establishing finger contact, subtle facial acting, or physical plausibility. Silent proxies cannot verify speech, knocks, or breathing. Such details remain target-generation instructions rather than claims about evidence present in the proxy.

Targeted correction and packaging. Corrections address specific observed defects and are followed by re-inspection (Fig. 8, step 5). For example, in *The key under the cup*, a hand in the intermediate proxy appeared too close to the cup during the reaction shot. A targeted repair instruction moved both hands away while preserving camera cuts, framing, role colours, and the cup and key trajectories. Later source and rendered-frame checks supported the cuts at 4 and 8 seconds. When the proxy agrees with the intended scene but an explicit prompt description conflicts with it, we revise the prompt to match the proxy.

We organize the corrected inputs into benchmark cases (Fig. 8, step 6). For scenes with all three proxy levels, the five core configurations are Pr_0T_0, Pr_0T_1, Pr_0T_2, Pr_1T_2, and Pr_2T_2, with additional variants where applicable. We also retain available Blender scenes, code, transcripts, metadata, previews, and earlier renderings to document the construction and revision process. The illustrated example has no image or audio conditions; the optional reference branches apply to other scenes.

A.6. Video Generation Settings

We evaluate Seedance-2.5, Kling-4.0, Wan-3.0, Seedance-2.0, and MiniMax-H3 using a common 720p resolution setting. Each scene specifies its target duration and framing, with 15-second videos and 16:9 or 21:9 aspect ratios as the standard formats. For each case, all supported models receive the same text prompt and reference conditions. The proxy is supplied directly as a reference video, with image, proxy, and audio inputs ordered consistently.

We retain one output per successfully generated model–case pair. Results are grouped by independent audio-reference availability, with Kling-4.0 excluded from the audio-reference group. Missing outputs receive zero, while

Table 10. The 32 criteria and representative evidence.

ID	Criterion	Evidence	ID	Criterion	Evidence
P1 — Spatial Structure			**C2 — Motion & Physical Plausibility**		
P1.1	Entity presence and count	Counts and tracks	C2.1	Kinematic and dynamic plausibility	Articulation; inertia
P1.2	Projected position and scale	Boxes; screen positions	C2.2	Temporal appearance coherence	Flicker; drift
P1.3	Relative spatial relations	Timed relation checks	C2.3	Contact and support plausibility	Contact intervals
P1.4	Occlusion and depth order	Occlusion; depth cues	C2.4	Dynamic geometry stability	Shape consistency
P2 — Motion & Camera Control			**C3 — Proxy Abstraction Handling**		
P2.1	Coarse entity trajectory	Proxy-aligned tracks	C3.1	Pseudo-action correction	Sliding; clipping; jumps
P2.2	Direction and interaction phase	Action phases	C3.2	Shape and structure completion	Target-object realization
P2.3	Camera motion and shot plan	Viewpoints; shot cuts	C3.3	Fine-grained motion completion	Missing articulation
P2.4	Framing and composition	Subject scale; framing	C3.4	Semantic detail completion	Specified semantic details
P3 — Temporal & Shot Continuity			**B1 — Entity & Appearance Grounding**		
P3.1	Event presence and order	Temporal event checks	B1.1	Text-to-entity attributes	Entity-level attributes
P3.2	Event timing and duration	Onsets; temporal windows	B1.2	Image-to-entity appearance	Reference matching
P3.3	Cross-shot action continuity	Action states across cuts	B1.3	Cross-shot or cross-view identity	Identity across views
P3.4	Persistent object state	Persistent state checks	B1.4	Negative binding and leakage	Target / non-target checks
C1 — Visual Realization Quality			**B2 — Action & Audio-event Grounding**		
C1.1	Realization completeness	Appearance of targets	B2.1	Actor–action–target binding	Actor and target roles
C1.2	Material, texture, and lighting	Material / light consistency	B2.2	Sound-event correctness	Required sound / source
C1.3	Perceptual and imaging quality	Image-quality cues	B2.3	Voice and source assignment	Speaker / timbre matching
C1.4	Local visual artifacts	Local distortions	B2.4	Audio–visual synchronization	Event onset alignment

Table 11. Checklist construction and verification.

Stage	Input evidence	Procedure	Output
Inventory	Text and optional references	Identify named entities, attributes, actions, and sound events; retain concrete referents.	Target requirements for checklist questions.
Proxy cues	Proxy frames and inventory	Distinguish *visible*, *abstract stand-in*, and *not observable*; unseen details are not automatically optional.	Structural evidence and completion targets.
Questions	Proxy, text, images/audio	Name the target entity or event; use text for intended content and the proxy for spatial and temporal structure.	Case-specific questions for P/C/B.
Screening	Proposed questions	Check answer form, polarity, grounding in available inputs, and unintended proxy-detail leakage.	Verified questions shared across models.
Video judgment	Questions and output video	Score with visual/audio evidence and the criterion rubric; subquestions do not receive separate weights.	Criterion scores for aggregation.

unsupported configurations and cases that cannot be evaluated are excluded from the corresponding evaluation set.

B. Evaluation Protocol

B.1. Checklist Construction

For each case, Gemini-3.1-Pro instantiates the applicable criteria in Table 10 into questions grounded in the proxy, text, and available reference media. Qwen3.5-397B-A17B checks whether each question is clear, supported by visible inputs, and aligned with the intended criterion. Invalid questions are revised before video scoring, with deterministic checks used to normalize question scope and answer format. The resulting checklist is fixed and shared across models for that case. Multiple evidence subquestions contribute to a single criterion-level assessment without receiving separate weights.

Question formulation. Checklist construction starts from an inventory of the named entities, attributes, actions, and events in the supplied conditions. The proxy contributes spatial and temporal observations, while text and optional reference media specify the intended appearance and assignments. Questions are written about concrete entities and localized events rather than generic references to an unspecified "action" or "object." Table 11 summarizes the information flow and verification requirements.

Proxy evidence. We construct a semantic inventory from the supplied text, then assign each requirement one of three

Table 12. Illustrative case-specific checklist covering all 32 criteria for a bakery dialogue scene. Questions are shortened and paraphrased for presentation rather than reproduced verbatim. Graded checks use the 0–4 scale; trajectory matching uses a categorical answer. The remaining checks use binary verdicts.

ID	Case-specific question / requirement	ID	Case-specific question / requirement
P1 — Spatial Structure		**C2 — Motion & Physical Plausibility**	
P1.1	Are one violinist, one baker, one parcel, one violin case, one table, two chairs, and the bakery fixtures present?	C2.1	Rate the naturalness of the reach and the characters' head and torso motions.
P1.2	Is the parcel in the lower-middle table area, occupying approximately 10–20% of the frame width as in the proxy?	C2.2	Rate temporal coherence of clothing and lighting across cuts.
P1.3	Are the violinist on the left, the baker on the right, and the parcel between them, with stable scene anchors?	C2.3	Does the baker's hand approach the parcel without clipping or implausible contact?
P1.4	Are tabletop occlusion and foreground shoulder depth order consistent with the proxy?	C2.4	Rate the stability of body proportions and hand geometry during the reach.
P2 — Motion & Camera Control		**C3 — Proxy Abstraction Handling**	
P2.1	Is the seated violinist's coarse trajectory the same, mostly the same, or different from the proxy?	C3.1	Rate the replacement of rigid proxy pivots and pseudo-actions with natural human motion.
P2.2	Does the baker reach toward the parcel and then retract, following the proxy's interaction phases?	C3.2	Rate completion of the characters, furniture, parcel, and other abstract proxy shapes.
P2.3	Are the static two-shot and two reverse over-the-shoulder views preserved in order?	C3.3	Rate fine articulation of faces, hands, and other motions omitted by the coarse proxy.
P2.4	Are the intended shot sizes and overall composition of the table, characters, and background retained?	C3.4	Rate the realization of clothing details, parcel wrapping, the violin case, and bakery fixtures.
P3 — Temporal & Shot Continuity		**B1 — Entity & Appearance Grounding**	
P3.1	Does the establishing conversation precede the parcel gesture and final reaction?	B1.1	Are the green and cream outfits assigned to the violinist and baker, respectively?
P3.2	Do the conversation (0–5 s), parcel gesture (5–10 s), and reaction (10–15 s) occur in their specified phases?	B1.2	Does each character match the facial appearance and hair of their assigned image reference?
P3.3	Does the reach/retract action remain continuous across corresponding cuts?	B1.3	Does the violinist retain the assigned identity and appearance across shots?
P3.4	Does the parcel remain intact, unopened, and stationary across all shots?	B1.4	Are the specified clothing attributes kept from leaking to the wrong character?
C1 — Visual Realization Quality		**B2 — Action & Audio-event Grounding**	
C1.1	Rate the completeness of the characters, props, and bakery setting, including natural faces and hands.	B2.1	Does the baker, rather than the violinist, gesture toward the wrapped parcel?
C1.2	Rate the plausibility of fabric, wrapping paper, wood materials, and soft window lighting.	B2.2	Are the specified spoken dialogue phrases audible?
C1.3	Rate focus, exposure, and freedom from imaging artifacts.	B2.3	Are the feminine and masculine voice references assigned to the violinist and baker, respectively?
C1.4	Rate rendering cleanliness, including the absence of clipping, melting, and corrupted textures.	B2.4	Are the characters' lip movements synchronized with their respective dialogue lines?

proxy-evidence states: *visible* when the relevant target is directly represented, *abstract stand-in* when a placeholder carries useful structure but not final appearance or articulation, and *not observable* when the proxy cannot establish that detail. These states prevent missing proxy evidence from being mistaken for permission to omit a text-specified entity or event. They also clarify why following a placeholder trajectory belongs to Preserve, whereas replacing unintended rigid articulation belongs to Complete.

Evaluator settings. Gemini-3.1-Pro serves as the primary visual and audio judge, using temperature 0. By default, the proxy and candidate timelines are presented side by side, together with uniformly sampled candidate frames (eight or twelve, depending on the capability), applicable reference images, and detail crops when needed. Requests use up to 64 visual inputs; audio-related judgments receive the candidate audio and relevant reference audio where required. The judge returns a structured answer, confidence, observed evidence, supporting timestamps, and criterion-specific subchecks. In the P/C/B evaluation pipeline, Qwen3.5-397B-A17B screens the checklist questions before video scoring.

Criterion boundaries. The categorization is determined by the *requirement being tested*, not solely by the visual phenomenon. Preserve measures proxy-specified structure; Complete measures realization of unspecified content and correction of unintended proxy abstraction; Bind measures assignment of text, image, and audio conditions to the designated entities and events. For example, P1.1 checks the

number of entities, whereas C1.1 checks their realized appearance. Similarly, P2 checks whether a coarse proxy trajectory is respected, C2 judges the naturalness of the motion, and B2 checks whether an action is assigned to the specified actor. This limits duplicate scoring of one defect under multiple labels.

Checklist validity. We audit input support at the level of instantiated assessments across evaluated outputs, rather than counting unique criterion definitions. Among 41,067 check instances, 50 were flagged for applicability review, and none remained unsupported after input-dependent filtering. This assessment addresses question grounding, complementing the human validation of video-level judgments in Sec. E.4.

Full checklist example. Table 12 illustrates how the 32 criteria are instantiated for a bakery dialogue scene with two seated characters, a wrapped parcel, a three-shot sequence, and image and voice references. Each question specifies the entities, events, or realization requirements to be assessed. Only applicable criteria enter the case-level evaluation.

B.2. Criterion Scoring and Applicability

Binary checks receive 1 if the expected answer is established and 0 otherwise. For P2.1, both *same* and *mostly same* count as a pass. Graded Complete checks use criterion-specific integer rubrics from 0 (not realized or severe defects) to 4 (fully realized with no relevant defects), normalized by dividing the level by 4. Compound binary requirements remain one score; a failed required subcheck overrides an affirmative overall verdict. Preserve, Bind, and the binary contact check C2.3 use a pass threshold of 0.50; the other Complete criteria use 0.60. Inapplicable criteria do not contribute to capability or dimension scores.

Cross-configuration evaluation. Matched configurations share a common criterion taxonomy, with questions instantiated from the available inputs. Preserve and Complete assess structural adherence and content realization, while additional image and audio references introduce grounding requirements under B1 and B2, respectively. Reference-specific binding is assessed on the corresponding applicable criteria.

B.3. Score Aggregation

Let $s_{i,k} \in [0,1]$ be the score for criterion k in case i, and let $\mathcal{K}_{i,g}$ be the applicable criteria for capability g. With the fixed criterion weights $a_k = 1$, the capability score is

$$f_{i,g} = \frac{\sum_{k\in\mathcal{K}_{i,g}} a_k s_{i,k}}{\sum_{k\in\mathcal{K}_{i,g}} a_k}. \tag{3}$$

For dimension d, let $\mathcal{G}_{i,d}$ denote its applicable capabilities. The dimension score averages these capability scores with equal weights:

$$D_{i,d} = \frac{1}{|\mathcal{G}_{i,d}|} \sum_{g\in\mathcal{G}_{i,d}} f_{i,g}. \tag{4}$$

The case-level mean and structure-sensitive scores are

$$Q_i = \frac{1}{|\mathcal{D}_i|} \sum_{d\in\mathcal{D}_i} D_{i,d}, \tag{5}$$

$$\rho_i = \frac{1}{12} \sum_{k\in\mathcal{P}} \mathbf{1}[\mathrm{pass}(i,k)], \tag{6}$$

$$S_i = Q_i \rho_i^2. \tag{7}$$

Here, $\mathcal{D}_i$ contains the applicable dimensions and $\mathcal{P} = \{\mathrm{P1.1}, \ldots, \mathrm{P3.4}\}$ is the fixed set of 12 critical Preserve criteria. The plans for all 294 proxy-conditioned cases support these 12 criteria, giving an effective denominator of 12 for scored proxy-conditioned outputs. The 18 no-proxy counterparts are evaluated separately using the control metrics below. *Joint Structural Success* is $\mathbf{1}[\rho_i = 1]$; it is not a joint pass requirement over P/C/B.

B.4. Control Metrics

To measure the benefits of structural guidance, we evaluate 50 matched proxy-conditioned and no-proxy video pairs from 18 scenes, with a shared target scene description, other references, and generation settings. We assess action transitions, cross-shot spatial consistency, and multi-view geometric consistency using complementary evaluation protocols.

Action Transition. We adopt the temporal state-transition and ordering checks of TC-Bench and the MLLM-based action-binding and object-interaction evaluation of T2V-CompBench, adapting them to our dynamic-action scenes. Each scene contains 2–4 predefined binary assertions covering action occurrence, state-transition completion, and temporal order. Qwen3.5-397B-A17B evaluates each generated video against the same assertions, returning a pass/fail judgment for each requirement.

Cross-shot Spatial Consistency. We adopt the relational evaluation of T2V-CompBench and the cross-shot continuity checks of MSVBench, adapting them to multi-shot scenes. For each scene, we define fixed assertions concerning persistent entities, landmarks, object states, and spatial relationships across shots. Using predefined shot boundaries, we sample one timestamped frame per shot. Qwen3.5-397B-A17B evaluates these frames in temporal order, allowing viewpoint changes while checking the spatial topology.

Table 13. Fidelity effects ($Pr_2 - Pr_0$) on 31 matched scenes.

Model	ΔP	ΔC	ΔB
Seedance-2.5	+0.022	-0.060	+0.040
Kling-4.0	+0.030	-0.072	-0.020
Wan-3.0	+0.011	-0.056	-0.036
Seedance-2.0	+0.016	-0.048	-0.089
MiniMax-H3	+0.067	-0.079	-0.182
Mean	**+0.029**	**-0.063**	**-0.057**

Assertion-based Scoring. For either metric m, let $\mathcal{A}_s^m$ denote the predefined assertions for scene s, and $J(v, a) \in \{0, 1\}$ indicate whether video v satisfies assertion a. The video-level score and paired improvement are

$$q_m(v, s) = \frac{1}{|\mathcal{A}_s^m|} \sum_{a \in \mathcal{A}_s^m} J(v, a),$$
$$\Delta_m = \frac{1}{N_m} \sum_{i=1}^{N_m} \left(q_m(v_i^+, s_i) - q_m(v_i^-, s_i) \right), \tag{8}$$

where v_i^+ and v_i^- denote proxy-conditioned and no-proxy outputs, respectively, and N_m is the number of valid matched pairs. Thus, assertions are averaged within each video before computing the mean paired difference.

Multi-view Consistency. We adopt the official MEt3R metric to evaluate geometric consistency across overlapping views. MEt3R compares corresponding scene content using estimated geometry and visual features without requiring ground-truth camera poses. We apply it to valid overlapping view pairs, retaining dynamic content, and aggregate the resulting errors. Unlike the assertion-based metrics, lower MEt3R scores indicate better consistency.

Pairwise VLM Preference. We additionally conduct holistic VLM comparisons on all 50 matched pairs. Each pair is evaluated in both presentation orders to mitigate positional bias. The primary order yields 40 proxy wins, two ties, and eight losses. Combining both orders produces 35 consistent proxy wins and 15 ties or order-dependent outcomes, with no consistent proxy losses.

The action-transition, cross-shot spatial, and MEt3R evaluations include 22 pairs from eight scenes, 14 pairs from five scenes, and 17 pairs from ten scenes, respectively. These metric-valid subsets may overlap, while VLM preference covers all 50 matched pairs. Together, the measures quantify complementary aspects of the structural benefits reported in main Table 2.

Table 14. Joint Preserve and Complete changes across 155 matched model–scene pairs.

	$\Delta C < 0$	$\Delta C \geq 0$	**Total**
$\Delta P > 0$	45	29	74
$\Delta P \leq 0$	49	32	81
Total	**94**	**61**	**155**

C. Proxy Fidelity

C.1. Matched Fidelity Trends

We analyze the effect of proxy fidelity on 31 matched scenes, with all three proxy levels (Pr_0, Pr_1, and Pr_2), fixed T_2 text, and no independent audio references. Table 13 reports the model-level changes from Pr_0 to Pr_2. All five models exhibit higher mean Preserve and lower mean Complete at Pr_2 than at Pr_0, while Bind responds differently across models. Notably, Seedance-2.5 gains 0.040 in Bind, whereas MiniMax-H3 loses 0.182, indicating that stronger proxy guidance does not necessarily facilitate reference grounding.

The intermediate fidelity level reveals a further distinction. Averaged equally across models, Preserve increases from 0.738 at Pr_0 to 0.776 at Pr_1, but slightly decreases to 0.767 at Pr_2. In contrast, Complete declines monotonically from 0.601 to 0.554 and 0.538. Most of the completion loss therefore occurs between Pr_0 and Pr_1 (-0.047), with a smaller additional decline from Pr_1 to Pr_2 (-0.016). The corresponding average scores $Q = (0.696, 0.668, 0.665)$ show that additional proxy detail does not translate into higher overall performance. These results suggest diminishing structural benefits alongside persistent completion costs, rather than a uniformly beneficial increase in proxy fidelity.

C.2. Scene-Level Variation

To examine whether the fidelity trade-off holds across individual scenes, we analyze 155 matched model–scene pairs from the same 31 scenes. Table 14 and Fig. 9 summarize the joint changes in Preserve and Complete from Pr_0 to Pr_2. Notably, 49 of 155 pairs (31.6%) exhibit reduced Complete without improved Preserve, indicating that additional proxy detail can incur completion costs without corresponding structural benefits. Conversely, 29 pairs (18.7%) improve Preserve without sacrificing Complete. Scene-cluster bootstrap estimates further show that the Complete decline ($\Delta C = -0.063$, 95% CI $[-0.082, -0.043]$) is more consistent than the Preserve improvement ($\Delta P = +0.029$, 95% CI $[-0.030, +0.095]$). These results highlight the importance of selectively exploiting proxy guidance rather than uniformly following additional proxy detail.

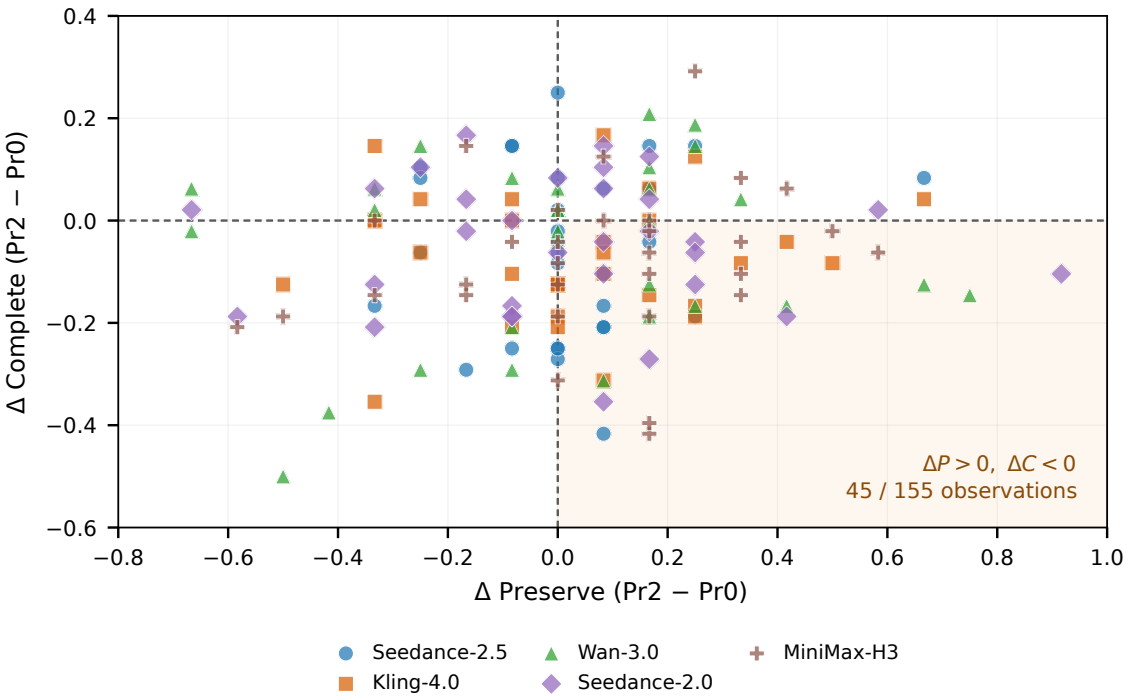


Figure 9. Scene-level Preserve and Complete changes from Pr_0 to Pr_2 (155 matched observations).

C.3. Completion Bottleneck

To identify which capabilities drive the preservation–completion tension, we further decompose the fidelity results into eight capabilities. Using the same 31 matched T_2 scenes and five models, Table 15 reports the model-averaged changes from Pr_0 to Pr_2, together with scene-cluster bootstrap confidence intervals and the direction of change across scenes.

The completion decline is dominated by C3 (Proxy Abstraction Handling), which decreases by 0.104 (95% CI: [-0.151, -0.057]) and shows negative changes in 24 of 31 scenes. Its decline is more than twice that of C2 (-0.048) and accounts for approximately 55% of the summed decreases across the three Complete capabilities. C1 also exhibits a smaller but consistent decline (-0.036), whereas the uncertainty interval for C2 overlaps zero.

In contrast, all three Preserve capabilities show positive mean changes but substantial scene-level variability, with confidence intervals spanning zero. This reveals an asymmetry in how models respond to increased proxy detail: structural adherence may improve, but the ability to replace proxy abstractions with completed appearance and motion becomes more challenging. The particularly large decline in C3 suggests that the main difficulty lies in distinguishing useful structural guidance from proxy-specific details that should be refined, rather than a uniform deterioration in visual or motion quality.

D. Condition and Binding Analysis

D.1. Input Richness and Binding

Structural proxies provide spatial and temporal guidance, while text and additional references specify the desired realization and the assignment of attributes and events. Increasing input richness therefore provides more precise guidance while introducing additional grounding requirements. Table 16 summarizes the configuration-level Strict differences, complemented by fine-grained binding diagnostics.

Text as complementary guidance. At fixed Pr_0, moving from T_0 to T_2 yields higher observed Strict scores for all five models, with gains of 2.4–7.7 percentage points. Detailed text supplies appearance, interaction, and temporal requirements beyond the coarse proxy structure. However, Seedance-2.0 achieves its highest text-level Strict at T_1, indicating that the benefit depends on the model rather than increasing monotonically with prompt detail.

Reference grounding. Reference-conditioned generation reveals a distinction between recognizing reference content and grounding it in the intended scene. Identity matching (0.897) and cross-shot identity consistency (0.909) exceed

Table 15. Capability-level fidelity changes ($Pr_2 - Pr_0$). All capabilities are evaluated on 31 matched scenes except B2 ($n = 21$). Signs count scene-level model-averaged changes; confidence intervals are obtained by scene-cluster bootstrap.

Capability	Mean Δ	95% CI	Scenes $+/-/0$	Evaluation focus
P1 Spatial Structure	+0.034	[-0.048, +0.113]	15/12/4	Entity placement and spatial relationships
P2 Motion & Camera	+0.024	[-0.050, +0.108]	15/10/6	Coarse trajectories and camera control
P3 Temporal Continuity	+0.029	[-0.044, +0.110]	16/13/2	Event order and cross-shot continuity
C1 Visual Realization	-0.036	[-0.073, -0.002]	12/17/2	Appearance, texture, and visual coherence
C2 Motion & Physics	-0.048	[-0.102, +0.012]	8/21/2	Fine-grained motion and physical interactions
C3 Proxy Abstraction	**-0.104**	**[-0.151, -0.057]**	**7/24/0**	**Artifact correction and detail completion**
B1 Entity & Appearance	-0.060	[-0.134, +0.005]	7/10/14	Entity identity and appearance grounding
B2 Action & Audio-event	+0.010	[-0.171, +0.200]	8/9/4	Action, voice, and event grounding

Table 16. Input richness and reference grounding. ΔStrict denotes model-averaged configuration-level differences in percentage points. Fine-grained binding scores are evaluated on their applicable subsets.

Contrast	Models	ΔStrict (pp)	Range (pp)	Fine-grained diagnostic
Pr_0: $T_0 \rightarrow T_2$	5	+5.2	+2.4 to +7.7	More detailed appearance, interactions, and event timing.
$Pr{+}T \rightarrow {+}I$	5	-10.4	-16.6 to -6.3	Identity: 0.897; cross-shot identity: 0.909; action–object: 0.636.
$Pr{+}T \rightarrow {+}A$	4	-13.8	-23.2 to -8.9	Voice/source: 0.768; audiovisual timing: 0.513.
$Pr{+}T \rightarrow {+}I{+}A$	4	-15.2	-24.0 to -4.4	Joint identity, appearance, and audio-event grounding.

action–object binding (0.636), while voice/source matching (0.768) exceeds audiovisual timing alignment (0.513). These correspond to gaps of 26.1 and 25.5 percentage points, respectively. Thus, reproducing reference appearance or sound does not necessarily ensure its correct assignment to the intended interaction or temporal event.

The lower configuration-level Strict scores under additional image or audio conditioning reflect the combined demands of structural preservation, content completion, and newly specified binding requirements. Together, these results highlight the challenge of coordinating multiple conditions within a fixed proxy-defined scene.

D.2. Targeted Cases

We examine 14 targeted cases through 12 paired comparisons involving reference removal, identity and voice assignment, event timing, and camera control (Table 17). The two voice-assignment comparisons each combine original and swapped configurations. These controlled variations reveal that structural preservation and condition binding can respond differently to changes in input requirements.

Identity swapping reduces Bind by 0.350 while Preserve decreases by only 0.033. Similarly, exchanging voice references in Dialogue A decreases Bind by 0.156 despite a 0.042 increase in Preserve, whereas Dialogue B maintains the same Bind score under both assignments. Camera-control variations also exhibit divergent responses: the dolly-only case improves Preserve by 0.200 while Bind decreases by 0.400. Reference removal produces heterogeneous configuration-level Bind changes, ranging from -0.344 to +0.417, as the available grounding requirements vary across scenes.

These paired comparisons demonstrate that preserving proxy-defined structure does not guarantee correct grounding of modified input conditions. The response to individual control changes also varies across scenes and requirement types, further motivating the separate evaluation of Preserve and Bind.

E. Statistical Validation

E.1. Statistical Protocol

We use scenes as the resampling unit to account for correlations among multiple configurations and model outputs derived from the same scene. Confidence intervals are estimated using 5,000 scene-cluster bootstrap replicates. Common-support comparisons retain the intersection of cases with scored outputs for every model within the corresponding input group, ensuring identical scene conditions and checklists. The common-support sets contain 190 cases from 48 scenes without independent audio and 44 cases from nine scenes with independent audio. Table 18 reports the corresponding results. These shared-case estimates complement the broader model comparisons in the main manuscript.

The common-support results reveal differences between average performance and structural reliability. Without independent audio, MiniMax-H3 achieves a higher Q than Seedance-2.0 (0.655 versus 0.636), but a lower Strict score (0.370 versus 0.395). With audio, Wan-3.0 achieves a higher

Table 17. Paired targeted-case comparisons. Original denotes the matched same-scene Pr_0–T_2 configuration, except for voice swaps, where it denotes the original speaker assignment. Scores are averaged over models with both outputs. Δ is Variant minus Original, calculated before rounding; N denotes the number of paired models (4 for reference-removal and voice cases, 5 otherwise).

	Preserve			Bind			
Variation	**Orig.**	**Var.**	ΔP	**Orig.**	**Var.**	ΔB	N
Joint image/audio removal A	0.625	0.688	+0.062	0.844	0.750	-0.094	4
Joint image/audio removal B	0.625	0.500	-0.125	0.594	0.250	-0.344	4
Joint image/audio removal C	0.583	0.771	+0.188	0.583	1.000	+0.417	4
Voice order swap A	0.625	0.667	+0.042	0.812	0.656	-0.156	4
Voice order swap B	0.854	0.833	-0.021	0.844	0.844	0.000	4
Identity swap	0.833	0.800	-0.033	0.850	0.500	-0.350	5
Image assignment swap	0.950	0.817	-0.133	0.925	0.900	-0.025	5
Delayed gate	0.783	0.767	-0.017	0.600	0.700	+0.100	5
Dolly only	0.750	0.950	+0.200	1.000	0.600	-0.400	5
Zoom only	0.750	0.867	+0.117	1.000	0.900	-0.100	5
Early focus	0.800	0.683	-0.117	1.000	0.900	-0.100	5
Centered camera	0.833	0.933	+0.100	0.800	1.000	+0.200	5

Table 18. Common-support performance with scene-cluster bootstrap 95% confidence intervals. Models within each input group are evaluated on identical cases.

Model	Q	**95% CI (Q)**	**Strict**	**95% CI (Strict)**
Without independent audio (190 cases, 48 scenes)				
Seedance-2.5	0.692	[0.663, 0.720]	0.464	[0.414, 0.514]
Kling-4.0	0.681	[0.650, 0.712]	0.442	[0.395, 0.487]
Wan-3.0	0.658	[0.626, 0.688]	0.415	[0.370, 0.459]
Seedance-2.0	0.636	[0.600, 0.671]	0.395	[0.339, 0.451]
MiniMax-H3	0.655	[0.624, 0.686]	0.370	[0.321, 0.419]
With independent audio (44 cases, 9 scenes)				
Seedance-2.5	0.659	[0.607, 0.707]	0.303	[0.232, 0.374]
Wan-3.0	0.651	[0.618, 0.686]	0.364	[0.297, 0.424]
Seedance-2.0	0.626	[0.576, 0.671]	0.293	[0.219, 0.370]
MiniMax-H3	0.601	[0.555, 0.644]	0.256	[0.212, 0.294]

Strict score than Seedance-2.5 (0.364 versus 0.303), despite a slightly lower Q. These differences illustrate how joint structural constraints can distinguish models beyond their average dimension scores.

E.2. Paired Comparisons

We further assess model differences using paired scene-cluster sign-flip tests on the common-support sets. Holm correction is applied across all model pairs within each input group. Table 19 reports representative comparisons with pointwise 95% confidence intervals and adjusted p-values.

The results indicate that some structural-performance differences are more robust than others. Without independent audio, Seedance-2.5 significantly outperforms MiniMax-H3 on Strict after Holm correction, whereas the gap between Seedance-2.5 and Kling-4.0 is not statistically significant. None of the audio-group pairwise differences survives correction. Thus, the benchmark reveals meaningful model-level differences while also distinguishing stable separations from closely matched performance.

E.3. Strict Sensitivity

To examine the influence of structural weighting, we vary the exponent in the Strict score while keeping the underlying criterion judgments fixed:

$$S_i^{(\alpha)} = Q_i \rho_i^{\alpha}, \qquad \alpha \in \{0, 1, 2, 3\}. \tag{9}$$

The primary definition uses $\alpha = 2$, while $\alpha = 0$ corresponds to the average dimension score Q. Table 20 reports results on the same input-group populations as the main benchmark.

The no-audio ranking remains stable across the tested exponents, apart from a tie at reported precision for $\alpha = 0$. In

Table 19. Selected paired Strict comparisons on common-support cases. Differences are computed before rounding. Confidence intervals are pointwise; p-values are Holm-adjusted across all model pairs in each group.

Model pair (first − second)	ΔS	**95% CI**	**Holm** p	**Finding**
Without independent audio				
Seedance-2.5 − Kling-4.0	+0.023	[-0.033, 0.076]	0.935	Top-two separation is uncertain.
Seedance-2.5 − Seedance-2.0	+0.069	[0.024, 0.113]	0.056	Positive difference, but not significant after correction.
Seedance-2.5 − MiniMax-H3	+0.095	[0.057, 0.133]	< 0.001	Significant structural performance gap.
With independent audio				
Seedance-2.5 − Wan-3.0	-0.061	[-0.146, 0.030]	0.980	No reliable separation.
Wan-3.0 − MiniMax-H3	+0.108	[0.035, 0.172]	0.199	Positive estimate, but not significant after correction.

Table 20. Sensitivity to structural weighting ($\alpha = 2$ is the primary setting).

Model	$\alpha = 0$	1	2	3
Without independent audio				
Seedance-2.5	0.678	0.538	**0.450**	0.387
Kling-4.0	0.663	0.517	**0.425**	0.362
Wan-3.0	0.644	0.488	**0.399**	0.340
Seedance-2.0	0.615	0.459	**0.375**	0.321
MiniMax-H3	0.615	0.439	**0.344**	0.283
With independent audio				
Seedance-2.5	0.666	0.446	**0.325**	0.251
Wan-3.0	0.584	0.416	**0.324**	0.265
Seedance-2.0	0.634	0.421	**0.314**	0.252
MiniMax-H3	0.581	0.350	**0.241**	0.179

contrast, the audio-group ranking changes as structural penalties become stronger: Wan-3.0 overtakes Seedance-2.5 at $\alpha = 3$. This indicates that Strict emphasizes a distinct aspect of model performance, while its effect on ranking depends on the distribution of joint structural successes. Reporting P/C/B alongside Strict therefore preserves both diagnostic detail and structure-sensitive comparison.

E.4. Human Validation

The main human evaluation covers 3,199 criterion-level judgments from 110 videos, with an overall agreement of 92.9% (89.7% for Preserve, 95.2% for Complete, and 94.8% for Bind). To further identify challenging evaluation capabilities, we examine 611 judge-labeled failures at the capability level (Table 21).

Human agreement on judge-labeled failures varies across capabilities. It is lowest for B2 (0.714), followed by P3 (0.830) and P1 (0.844), whereas B1 achieves 0.885. The results suggest that assessing event timing, sound-source assignments, and actor–action relationships is more challenging than assessing entity appearance, highlighting the importance of localized temporal and relational evidence for reliable evaluation.

Table 21. Capability-level human agreement on judge-labeled failures, with Wilson 95% confidence intervals.

Capability	**Fail** N	**Agreement**	**95% CI**
P1 Spatial structure	77	0.844	[0.747, 0.909]
P2 Motion / camera	46	0.935	[0.825, 0.978]
P3 Temporal continuity	47	0.830	[0.699, 0.911]
C1 Visual realization	71	0.986	[0.924, 0.998]
C2 Motion / physics	120	0.858	[0.785, 0.910]
C3 Proxy abstraction	119	0.933	[0.873, 0.966]
B1 Entity / appearance	96	0.885	[0.806, 0.935]
B2 Action / audio event	**35**	**0.714**	**[0.549, 0.837]**

F. Limitations and Future Work

Limitations. Our proxy fidelity levels jointly vary geometric and motion detail, making it difficult to isolate their individual effects. Moreover, evaluating subtle physical interactions and precise audiovisual synchronization remains challenging for current VLM-based evaluators.

Future Work. Our findings motivate more effective proxy-guided generation methods that selectively preserve intended structure while refining incomplete geometry and motion. Future research could explore adaptive proxy conditioning and proxy-aware training to better balance Preserve and Complete, alongside stronger entity- and event-level multimodal grounding to improve Bind.